Human raters *and what sways them*

What published studies show about human ratings of model output: what raters respond to, why studies resist comparison, and where a large public vote is disputed.

MITWA LAB, MUMBAI

IN PLAIN WORDS

When a chatbot's answers are judged, the judges are often people, and their ratings are treated as the standard other measures are checked against. This piece puts five questions about those ratings to published studies. In one study of older models, crowd raters could not reliably tell GPT-3 text from human text, and in another they rated GPT-2 stories close to human-written ones on some days. A single overall score is associated with how confident, or how long, an answer sounds, though the studies cannot always tell bias from real quality. Rating studies are described so thinly that they can seldom be compared or repeated. A very large public vote can still be skewed, and the people who run one dispute how far. A company's approval tests missed a flaw. The studies suggest that ratings from people need checking too.

Why a person's rating is treated as the answer key

An open-ended answer has no list of correct answers to mark it against, so the usual stand-in is a person, or many people, asked which output is better or how good it is. Hosking et al. (2024) call this single overall score the de facto standard for open-ended generation, and Belz et al. (2023) note that human evaluation is widely treated as the litmus test of quality in the field. Belz et al. add that automatic measures are themselves meta-evaluated against human evaluation.

Our technical note 02, Eleven words, argues that scores from model judges mix quality with length, and it asks for ratings from people on at least a sample. This review asks what ratings from people bring with them by putting five questions to the studies below.

Can people tell good writing from bad, or even machine from human?

Clark et al. (2021) recruited untrained evaluators on Amazon Mechanical Turk, 130 for each of six settings (stories, news articles and recipes, each with text from GPT-2 or GPT-3 mixed with human-written text). When the choice was between GPT-3 and a human author, evaluators named the machine correctly 49.9% of the time against a chance level of 50%. For GPT-2 the figure was 57.9%. Three short training methods raised accuracy to at most about 55%, and the one significant overall gain was not significant within any single domain.

Karpinska et al. (2021) examined story fragments written by a GPT-2 model and by people, using 200 story premises and a five-point scale. Crowd workers with strict qualifications did not consistently rate the model's stories significantly lower than the human-written ones when each set was rated in its own task. On coherence the model averaged 3.82 and the human-written stories 4.11, 3.92 and 4.05 on three different days, and the authors say that, depending on the day used for comparison, the model is rated similarly to the human-written stories on all four properties. Three English teachers rated the human-written stories clearly higher on coherence (4.38 against 3.73) and relevance (3.82 against 2.54). Crowd workers shown both kinds of story side by side did score the model lower, which the authors read as a calibration problem.

Spotting who wrote a text is a different task from rating how good it is, and a reader can fail at the first while rating the second sensibly. Both studies used GPT-2, and Clark et al. also GPT-3, models described in papers from 2019 and 2020. Within those limits, the untrained evaluators in Clark et al. could not

reliably tell GPT-3 text from human text, and short training lifted their accuracy to about 55% at most. The strictly qualified crowd workers in Karpinska et al. rated GPT-2 stories close to human-written ones on some days, while three English teachers did separate the two kinds of story, though that rests on three people.

Does a single preference score measure quality?

Hosking et al. (2024) asked one group of crowdworkers to mark 900 outputs for ten kinds of error, and a separate group to give each output an overall score from 1 to 5. The outputs were summaries, product descriptions and how-to answers, from four models and from human-written references. Refusing to answer weighed most heavily in the overall score, while factual errors and inconsistency with the request weighed much less. Six of the ten error types contributed to the score, and the authors describe its coverage as fairly good, but they say a single score is likely to obscure failures on those two criteria.

In a second experiment the authors instructed models to answer more assertively or more cautiously, which produced 1,500 outputs, and compared the crowd's error marks with their own careful annotation of 300 examples per error type. Crowdworkers missed more factual errors than the authors did, and the shortfall grew when the text was assertive and shrank when it was cautious. Perceived assertiveness correlated with the overall quality score at 0.68. The authors say the direction of causation is unclear, since assertive answers might simply be better, and that their own annotations are not strictly unbiased.

In the second of two experiments in Steyvers et al. (2025), 179 participants (301 took part across both) estimated, from an explanation alone, the probability that a model's answer was correct. Longer explanations raised their confidence without improving their ability to separate right answers from wrong ones, although that task was judging correctness and not stating a preference.

The platform behind a very large public vote has examined the same question in its own data. Li et al. (2024), from the Chatbot Arena team, added answer length and counts of markdown headers, bold text and list items as extra terms in the model that turns votes into rankings. Rankings shifted, and length was the dominant style factor, with a fitted coefficient of 0.249 against 0.019 to 0.031 for the markdown features. The authors state that the analysis is observational and that length may itself track substance. A long answer winning is therefore not by itself evidence of bias, which is the strongest form of the opposing view, although the authors themselves read the ranking shift as an effect of style.

Ethayarajh and Jurafsky (2022) argue from utility theory that Likert scales can reverse the direction of a true preference in some cases. In their test, 90 crowdworkers compared stories from four sizes of GPT-3 and a human-written continuation. The usual average of ratings recovered two of five expected preferences, while their own protocol, which asks annotators to estimate how likely they are to prefer one system over another, recovered all five. This is one study, and three of the five expected preferences for the stories assume that larger GPT-3 variants write better text while the other two rest on Clark et al. (2021).

Can one rating study be set beside another?

Howcroft et al. (2020) annotated the 165 papers from two language-generation venues (2000 to 2019) that reported a human evaluation, covering 478 individual evaluations. The papers used 204 different names for the quality being rated. After sorting them by what was actually asked, the authors counted 71 distinct criteria, and found that fluency, one of the most familiar names, corresponded to 15 different ones. Definitions were missing for 279 of the 478 evaluations.

Belz et al. (2023) tried to organise a coordinated reproduction of such studies. Of 177 papers from two ACL venues between 2018 and 2022 that mentioned human evaluation with participants, 116 passed a first suitability check, and about 13% of those 116 (15 papers, by the paper's flow diagram) had both low enough barriers and enough obtainable detail to be considered for reproduction. Of the experiments the team then selected, all but one proved to have flaws that made a reproduction of doubtful meaning.

Card et al. (2020) looked at the statistics. In 117 Likert-scale comparisons from 41 papers at one 2019 conference, 69% used 100 or fewer items, and three annotations per item was the most common design (57%). Using variance estimated from public rating datasets, their simulation found that three raters on 100 items would be underpowered unless the true difference was about 0.2 or more on a scale rescaled to run from 0 to 1, in the higher-variance setting that fitted most of those datasets. That is a simulation under stated assumptions and not a measurement of any particular study. All three studies concern work

from before 2023, and practice may since have improved, which none of them measured.

Is a very large public vote neutral?

Chiang et al. (2024) set out the case for large public voting. Chatbot Arena records which of two anonymous answers a visitor prefers. By January 2024 it had over 240,000 votes from about 90,000 users. To check vote quality, graduate students fact-checked and labelled 160 battles in two model pairings. Agreement between the crowd and each of the two experts ranged from 72.8% to 83.1%, against 79.4% and 89.8% between the experts in the two pairings. The authors attribute the gap mainly to crowd voters overlooking factual errors, the weakness Hosking et al. (2024) measured in a different setting. They add that their users are probably mostly hobbyists and researchers.

Singh et al. (2025) audited about two million battles across 243 models from 42 providers (January 2024 to April 2025 by the paper's account, with some data from April 2023). They report that some providers could test several private variants of a model and publish only the best, with Meta testing 27 in the lead-up to a single release. They argue, and show by simulation, that choosing the best of several noisy scores biases the published score upwards. They also estimate that training a model on arena data raised its relative performance on Arena-Hard, a test drawn from the arena's distribution, by up to 112%, with limited gains on other tasks.

The platform's team replied in May 2025 (Arena Team, 2025). It says its policy on testing unreleased models was published in March 2024, that the plotted simulation used a variance the authors chose, and that its own analysis puts the lift from pre-release testing at about 11 points after 50 tests and 3,000 votes, shrinking as fresh votes arrive. It says the 112% result came from a static 500-item benchmark scored by a model judge, which is a different thing from human votes. It also says the two identical checkpoints the authors submitted received scores of 1069 and 1054 with overlapping intervals (the paper reports 1052 and 1069).

What is settled is narrower than either side's framing. The intervals the paper reports for the identical checkpoints overlap, so that single comparison does not by itself demonstrate bias. The paper's case rests on the selection mechanism. Singh et al. estimate about 17.7 Arena points for 50 identical variants at 3,000 votes, and say the inflation grows, and no longer vanishes as votes accumulate, when the variants truly differ. Their one live test with different variants put two versions of one model at 1097 and 1059 with nine models between them, a gap that mixes any real difference between the variants with the effect of selection. The platform's figure is about 11 points, shrinking to zero as fresh votes arrive. We have not re-run either analysis. Apart from any gaming, Li et al. (2024) report that controlling for style changes the rankings, which they read as an effect of style, so a vote can be honest and still not neutral about what it measures.

Does user approval settle whether a model is good?

In April 2025 OpenAI released a ChatGPT update to GPT-4o that it later described as noticeably more sycophantic, meaning that it aimed to please the user, and it began rolling the update back three days later. In the company's own account (OpenAI, 2025), offline evaluations generally looked good and a small A/B test suggested that the users who tried the model liked it. Some internal expert testers said its behaviour felt slightly off, and the company launched on the strength of the positive user signals, which it now calls the wrong call. It also reports that the update added a reward signal from thumbs-up and thumbs-down data, and that such feedback can favour more agreeable answers.

This is one company's retrospective account of one incident, and none of the sources we opened checks it from outside. On the company's own account, it suggests that approval signals can look positive while a flaw is present. Human judgement did register that something was off, from testers whose impressions differed from the A/B signal, although sycophancy was not explicitly flagged by them. One part of the company's stated response is to give the testers' qualitative signals more weight in launch decisions.

What follows from these studies

What follows is our argument, and it is not a result from any of the studies. A human rating is a measurement, and like any measurement it carries an error that depends on who rates, what they are asked, what they are shown and how many of them there are. Several of the errors reported here lean the same way. Crowd raters missed more factual errors in assertive text (Hosking et al., 2024), longer explanations raised confidence without improving discrimination (Steyvers et al., 2025), and length was the dominant style effect in the Arena's own votes (Li et al., 2024), although the last may partly reflect real quality. Technical note 02 reports, from the literature, a lean towards length in model judges. When two

kinds of judge share a bias, agreement between them is weak evidence that either is right.

A rating protocol can itself be tested, by asking whether it recovers differences expected to exist, as Ethayarajh and Jurafsky (2022) did, or whether untrained and trained raters reach the same judgements, as in the comparison with teachers by Karpinska et al. (2021). This review argues that a rating from people, like a score from a model, is a measurement whose error can be estimated. A claim resting on a rating is easier to weigh when it says how the rating was made.

What is not yet known

- Whether the failures of crowd raters on GPT-2 and GPT-3 text hold for ratings of current models by current raters. None of the sources we opened repeats the authorship-detection test on current models. Ethayarajh and Jurafsky (2022), using a different protocol on GPT-3 stories, found the human writer significantly less preferred to davinci under the standard protocol, a result they call surprising and likely incorrect, which differs from Clark et al. (2021).
- How people rate emotionally sensitive conversation. The outputs in these studies were summaries, product descriptions, stories, recipes, news articles and general chat answers, and none of the sources we opened reports a rating of a supportive exchange over many turns. Chatbot Arena votes include multi-turn conversations, but the sources do not say whether any are supportive.
- How large the effect of private selection is on the live leaderboard, where Singh et al. and the platform give different estimates.
- Whether length and assertiveness cause higher ratings or merely accompany better answers. Li et al. (2024) describe their analysis as observational, and none of the sources we opened varies length or assertiveness independently and then measures overall quality ratings, though Steyvers et al. varied explanation length and measured confidence in correctness.
- Whether trained raters remove the assertiveness skew. The reference annotations in Hosking et al. (2024) were the authors' own.

References

Arena Team (2025). Our response to "The Leaderboard Illusion" writeup. Arena (formerly LMArena), news post, published 9 May 2025, last updated 13 June 2025. news.lmarena.ai/our-response

Belz, A., Thomson, C., Reiter, E., Abercrombie, G., Alonso-Moral, J. M., Arvan, M., et al. (2023). Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. arXiv:2305.01633 (v2, 7 August 2023). arxiv.org/abs/2305.01633

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. (2020). With little power comes great responsibility. arXiv:2010.06595 (v1, 13 October 2020). arxiv.org/abs/2010.06595

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. arXiv:2403.04132 (v1, 7 March 2024). arxiv.org/abs/2403.04132

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. arXiv:2107.00061 (v2, 7 July 2021). arxiv.org/abs/2107.00061

Ethayarajh, K., and Jurafsky, D. (2022). The authenticity gap in human evaluation. arXiv:2205.11930 (v3). arxiv.org/abs/2205.11930

Hosking, T., Blunsom, P., and Bartolo, M. (2024). Human feedback is not gold standard. International Conference on Learning Representations (ICLR 2024). arXiv:2309.16349. arxiv.org/abs/2309.16349

Howcroft, D. M., Belz, A., Clinciu, M., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020), 169-182. aclanthology.org/2020.inlg-1.23/

Karpinska, M., Akoury, N., and Iyyer, M. (2021). The perils of using Mechanical Turk to evaluate open-ended text generation. arXiv:2109.06835 (v1, 14 September 2021). arxiv.org/abs/2109.06835

Li, T., Angelopoulos, A., and Chiang, W.-L. (2024). Does style matter? Disentangling style and substance in Chatbot Arena. LMSYS Org blog, 29 August 2024. lmsys.org/blog/2024-08-28-style-control

OpenAI (2025). Expanding on what we missed with sycophancy. Company post, 2 May 2025. openai.com/index/expanding-on-sycophancy/

Singh, S., Nan, Y., Wang, A., D'souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., Ermis, B., Fadaee, M., and Hooker, S. (2025). The leaderboard illusion. arXiv:2504.20879 (v2, 12 May 2025). arxiv.org/abs/2504.20879

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L., and Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7(2), 221-231. doi.org/10.1038/s42256-024-00976-7 (read as arXiv:2401.13835v2, arxiv.org/abs/2401.13835)